DOCUMENT RESUME

ED 204 375 TM 810 363

AUTHOR Elliott, John
TITLE Evaluation Measures and Methods: Some Intersections.
PUB DATE [80]
NOTE 15p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluation Methods; *Innovation; *Literature Reviews; *Measures (Individuals)
IDENTIFIERS *Traditionalism

ABSTRACT

The literature is reviewed for four combinations of evaluation measures and methods: traditional methods with traditional measures (T-Meth/T-Mea), nontraditional methods with traditional measures (N-Meth/T-Mea), traditional measures with nontraditional measures (T-Meth/N-Mea), and nontraditional methods with nontraditional measures (N-Meth/N-Mea). What is traditionally measured is cognitive achievement before and after a treatment (i.e. the program). The method used is traditionally the administration of an achievement test whose score is afterwards statistically manipulated. The approach is research-oriented in that an hypothesis is made before the evaluation begins. It is frequently difficult to separate what is measured ("measures") from how it is being measured (method). The N-Meth/N-Mea category contained the largest number of examples in the literature and may be described as one in which materials other than tests and direct observation are examined by other than statistical methods. "Triangulation" has been defined as the systematic use of different methods to cut down on validity threats. The use of a variety of techniques increases reliability by cross-checking. Triangulation seems to combine the advantages of both the traditional and nontraditional approach. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204375

EVALUATION MEASURES AND METHODS: SOME INTERSECTIONS

by John Elliott

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Elliott

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 810 363

Evaluation Measures and Methods: Some Intersections

Joel Weiss (1980), has suggested that when we combine methods of evaluation with the types of things that could be measured, we might have four combinations: traditional methods with traditional measures, non-traditional methods with traditional measures, traditional methods with nontraditional measures, and nontraditional methods with nontraditional measures. Graphically, the following grid might show the possibilities:

	Traditional Methods	Nontraditional Methods
Traditional Measures		
Nontraditional Measures		

The literature will be described for each quadrant. The summary will comment on the quality and quantity of the references for each intersection.

Traditional Measures, Traditional Methods

One difficulty in using these labels is in deciding what is traditional. Ralph Tyler says it is the pre-test post-test model, which he finds discriminates among students, rather than competing curricula (1967). This type refers to a method, but not a measure. Achievement tests are conventional measures that can be used in a pre-test post-test manner. When used in this way they don't discriminate among students on a logical basis, (West, 1977), or describe interactions between the teacher and student, (Palonsky, 1977). These measures assume that validity depends on repeatability or ability to be replicated, (Pilsworthy and Ruddock, 1975).

Atkinson, (1979), and Glaser and Strauss, (1967), call normal evaluation a hypothetical-deductive method in which a guess or intuition is tested. They call it the research approach. Cronbach, (in Mason, 1973), uses different kinds of measures for different purposes, such as selection and classification of persons, evaluation of treatments, and checking on scientific hypotheses. He notes that practitioners as teachers use test information for diagnostic and prescriptive purposes, administrators for program and allocation of resources.

Researchers are interested in relationships among variables, and making causal inferences, while developers use both approaches.

The "Black box" is most often described as the traditional approach, (Atkinson, 1979, West, 1977, Magoon, 1977, Mason, 1973, Baird, 1973.) This is the attempt to link instructional methods to student achievement without looking at the intervening variables.

Statistical manipulations are commonly associated with traditional measures and methods. Parlett and Hamilton claim that the statistical investigator has the alternatives of randomizing using large numbers, or strictly controlling. West states that single factors such as socio-economic status are used in statistical studies when interacting variables are needed. To be fair, he seems to be ignoring multi-factor analysis. He suggests that the quantitative approach aims at a general understanding, and may not apply in particular cases.

So that the statistics will involve simple proportions, in traditional evaluation dichotomous scoring of items takes place (i.e. that the answers are either right or wrong), (Wells and Hively, 1973).

It is likely also for statistical reasons that the cognitive domain is stressed in most evaluations. Weiss finds that in major studies, affect is either not measured, or the results are ignored by policy makers.

Formative and summative tests do not, according to Mehlinger, (in Schaffarick and Hampson, 1975), separate along traditional and nontraditional lines. The results of formative testing are normally assigned to developers, and of summative testing to consumers.

Values are considered in curriculum evaluation. Vallance (1977), makes the point that the only modes used are technical, political, and scientific.

In describing the i.t.a. evaluation, Hamilton, (1976), summarized the main aspects of, and complaints about traditional evaluation. The typical comparative or preordinate study has problems of novelty, consciousness of being experimented on, and difficulty of matching experimental and control

groups. This goal-based instruction ignores unintended outcomes, imposes on the receiver of the information a philosophical point of view, and does not compare the program with any other approach. Its statistics ignore the processes that go on in the classroom. The objectives testing emphasizes what is easily tested. Redevelopment in midstream is discouraged. The criteria of the practitioners are ignored. Scriven's unintended outcomes are ignored.

Summary: Traditional Measures, Traditional Methods

What is normally measured is cognitive achievement before and after a treatment, (i.e. the program). The method used is usually the administration of an achievement test whose score is afterwards statistically manipulated.

The approach is also research-oriented in that an hypothesis is made before the evaluation begins.

As could be noted with all four intersections, it is frequently difficult to separate what is measured, (here called "measures"), from how it is being measured, (method).

Traditional Measures, Nontraditional Methods

If the score on a written post-test is allowed to represent cognitive ability, then it might be claimed that a traditional measure was used. Combined with this could be a variety of testing or group sorting methods.

Time is one factor that needs to be considered when nontraditional methods are proposed. Hampson cite Macdonald that objectives can only be known after an act of instruction. Evaluation would then have to be summative. Hampson points out though, that the differences between formative and summative evaluation are differences of use, not form or time. He at the same time encourages learner, or post-publication evaluation.

EPIE, which also favours Learner Verification and Revision, recommends continuous, large-scale evaluation at the summative stage which might reveal aspects not clear in the formative levels.

The question of the use or relative timing of objectives to evaluation is also raised by Stake, who calls the objective in preordinate evaluation the stimulus and evaluation the response. Magoon takes a similar position.

He parallels preordinate evaluation and the associationist philosophical position in which the subject's behavior is a response to stimuli. Purvis, (in Schaffarick and Hampson), states that starting a curriculum with objectives is a bit like attempting to attach a dog to its tail.

Another method variable that sometimes includes time considerations, is the selection of the group to be studied. Smith and Carpenter (1972), begin a study with one teacher and then expand for context seeking similarities and differences. Parlett and Hamilton progressively focus on increasingly wider areas of the curriculum. All data are coded then checked by outside observers.

Glaser and Strauss use traditional tests, but select their groups based on whether they contain similar categories or properties, not on whether the groups are comparable. This might be similar to Anderson, et al's use of "marker variables", to be used in separate studies so that they could be compared. Mason also suggests that surveys have good variables that could be shared among, and allow comparisons across surveys.

Glaser and Strauss' selection of groups gives control over conceptual level, population scope, and provides simultaneous maximization or minimization of differences and similarities of data. Minimum differences shows the conditions under which the category exists, while maximizing provides a variety of information about the category itself.

Atkinson (1979), selects cases on the basis of their unusualness or extremity when compared to the norm. This "deviant", or "critical case analysis", has resemblances to Pilsworth and Ruddock's (1975), "critical incident method", in which events are only visible during a "moment of truth". In both cases, results, rather than other characteristics dictate selection. Rather than depending on change through random selection, the evaluator has used deliberate selection.

Rasch (in Pace), uses a conventional test, but only attends to those responses in which when two items are compared on a test, both a correct and an incorrect response are received. He speaks negatively of the adversary model,

in which a material is intentionally examined to discover what is wrong with it. LVR does examine for difficulties in a curriculum, so that a subsequent edition will be better. Parlett and Hamilton have an evaluator on the site play the role of a "devil's advocate". In either case, standard tests could be used to determine what was or was not wrong with a program.

Merlinger, (in Schaffarick), evaluates at all stages, from judging the rationale of the project, the statement of goals and assumptions about the competency of the learner to the evaluation of materials as they are marketed to plan future revisions.

Like Merlinger, Cronbach (in Mason), evaluates for different purposes. For different purposes, he measures differently. Methods would differ for selection and classification of persons, evaluation of treatments, and checking on scientific hypotheses.

Both traditional and nontraditional methods are used by Smith and Pohland, who gather qualitative data which lead to propositions which can be checked mathematically. Multiple settings are used to create a multi-method, multi-trait approach to ensuring validity. Depth is provided by the sheer mass of data. They and several others, (Russell et al, Glaser and Strauss, West, Atkinson, Parlett and Hamilton, Pilsworth and Ruddock) recommend the use of a variety of techniques, to increase reliability by cross-checking. Some of the methods are traditional, others not.

Summary, Traditional Measures, Nontraditional Methods

It is sometimes, but not always, nontraditional to use formative evaluation. Summative evaluation that allows revision of future editions of a material is unusual.

Some, but not all, of the authors prefer that the innovative method allow the objectives to emerge from the material.

Several evaluators avoid the randomization of statistics and instead select the groups and members of groups based on particular criteria.

Many authors use a combination of traditional and nontraditional methods, depending upon the stage of evaluation, and purpose. In all cases though, it is at least possible that the measure, (i.e. what is measured), would be traditional.

Nontraditional Measures, Traditional Methods

This cell would contain any situations in which measurement method is or could be traditional, (such as with pencil and paper tests), while the object of the evaluation would be untypical.

Time variables must be considered. Tyler wishes to search the literature before the project is begun. He also checks the implementation to be sure the new procedure is being used. Resnick also supports the study of implementation data to interpret outcome data.

Still within the time dimension is the area of unintended outcomes, (Purvis, Clark and Scriven, Jenkins, Atkinson, Stake, Magoon, Smith and Pohland). This responsive evaluation examines issues as they emerge from the data.

Some approaches are different from the norm by virtue of the fact that they evaluate unusual subjects. West expands the group to be studied to include teachers, pupils, janitors, principals, parents and administrators. Atkinson has us look at unusual or extreme cases to contrast with the normal.

Merlinger has perhaps the largest catalogue of nontraditional measures. He judges the rationale of the project, the statement of goals and assumptions about the competency of the learner, and the quality of the materials as they are marketed to plan future revisions. This last concern is also shared by EPIE, (LVR), and Lewy.

Glaser and Strauss equate more traditional measures such as library work, direct observation, questionnaires, and interviews with letters, biographies, autobiographies, memoirs, speeches, novels and jokes.

Summary, Nontraditional Measures, Traditional Methods

What may be a traditional measure at one stage of evaluation, may be nontraditional at another stage. Outcomes that have not been anticipated must be attended to. Everyone even remotely involved with the enterprise could be involved in the evaluation. Questions about rationale and quality can be asked. Unobtrusive cultural measures can be revealing.

Nontraditional Measures, Nontraditional Methods

Infrequently do writers specify both their measure and method. We might describe this category as one in which materials other than tests and direct observation are examined by other than statistical methods.

At times, it will be necessary to link one author's method to another's measure, if the combination seems possible.

Weiss would use affective tests only at the classroom level, in a formative way that will allow for examination of mediating effects (and perhaps affects). Messick gives varieties of noncognition, most of which are nonconventional measures:

1) experiential background factors (work, age, education, social class, etc. which equal proxy factors.
2) affects (positive and negative feeling).
3) attitudes (positive or negative feeling plus evaluation).
4) belief
5) interests
6) motive
7) temperament
8) social sensitivity and interpersonal competence.
9) cognitive style (e.g. scanning vs. focussing).
10) creativity
11) values.

He notes that there are no correct answers in a noncognitive test; we draw evidence from the consistency, not the correctness of the response.

While Pinar uses as his four structural elements, curriculum, place of the curriculum in history, response of the student, and the psychological and

biological context of the student, his explanations of how he evaluates these measures are not entirely clear. Willis, however, emphasizes student response and uses adjectives in his method.

Kelley studies the rhetoric of the curriculum: metaphor, point of view, plot, and theme. His method is a literary approach.

Warren assesses student purpose and preferences using questionnaires, interviews, group meetings with students, trial programs to test student responses, faculty-student retreates, student-run surveys, and evaluations, and direct observation of student behavior.

Smith and Carpenter (1972), begin a study with one teacher and then expand for context, seeking similarities and differences.

Wells et al (1973), developed "domain-referenced curricula", by modifying one objective or its contexts through changing its constituent parts. A domain of similar behavioral objectives is then created. What might be changed would be forms of instruction, of responses, syntax (the same response for different contexts for the stimuli words), vocabulary, sequence, supplementary information, (to add if the student fails to respond), numerical values, physical characteristics of materials, production, choice or judgment. If we had accounted for these, we might more often find differences in the results of gross treatments.

They would allow informal intrinsic evaluation only at the early stages, when judgments only were possible, since students were not yet using the material. Formal pay-off evaluation would take place later.

This simultaneous, or sequential use of traditional and nontraditional methods and measures is sometimes referred to as "triangulation".

Triangulation

Both West and Atkinson define this term as the systematic use of different methods to cut down on validity threats. Russell et al (1969), used scale measures and standardized tests. They also employed questionnaire responses and participant observation reports. These "three legs" stayed in

communication and shared findings in a formative way.

While Pilsworth and Ruddock use a variety of approaches, they especially stress the qualitative and phenomenological method. Their method is "illuminative", using historical, anthropological and biographical analyses.

The use of a variety of techniques increases reliability by cross-checking, as used by Parlett and Hamilton. Direct observations are used to build up a continuous record of ongoing events, transactions and informed remarks, with added interpretive comments. Codified observations of teacher-pupil interactions, interviews, documentary records of meetings leading up to innovations and eye witness accounts of an event contribute to the multidimensionality of the evaluation.

Glaser and Strauss do not use the term, but do use the method. As was mentioned in the nontraditional measures, traditional methods section, they give letters, biographies, autobiographies, memoirs, speeches, novels and jokes the same status as library work, direct observation, questionnaires and interviews.

Summary, Nontraditional Measures, Nontraditional Methods

Some approaches, such as Kelley's literary critic evaluation, combine method and measure. Warren's assessment of student purpose and preference also combines both.

Pinar's study of student response, requires such methods as Willis' use of adjectives.

"Triangulation", seems to combine the advantages of both the traditional and nontraditional approach. Most important facets are in this way covered, while the diversity of methods provides a cross-check.

Summary, Measures and Methods, Traditional and Nontraditional

There seems to be little agreement as to what is traditional and what is not in both methods and measures. Measurement by objectives is described

by both Stake and Magoon as traditional, and of questionable use. Cronbach agrees that this research approach is appropriate, but only for researchers, not developers. Vallance makes the point that while research says what works, evaluation tells how.

Perhaps the statistical manipulation of pretests and post-tests in standardized form would represent the traditional, while almost anything else would be nontraditional.

Most of the reports on evaluation in educational and psychological journals are traditional in both aspects. The references contained here were found by intentionally looking for nonconventioanl citations.

It was possible to find examples of all four types of intersections. The largest number of cases was found in the nonconventional methods - non-conventional measures category. This may partly have been because included in this group were those cases where a number of combinations were used, some conventional, some not, as with the "triangulation" group.

This group, and others have argued that the combination of measure and method that is appropriate depends upon the evaluation purpose. If this is true, then the quadrants here used cannot be compared on a good-bad, or even a useful-non useful basis, but on an appropriate-not appropriate continuum for a given audience or purpose.

A final caution is necessary. Someone has said that I.Q. is what is measured by an I.Q. test. If we accept this, then we must also believe that some affect methods define what we measure when we evaluate affect. Many of our nonconventional tests may have poor validity when used to measureboth conventional and nonconventional measures.

It does appear though that old wine can be used in new bottles, (and new wine in old bottles), at least in the area of evaluation.

Bibliography

Atkinson, Paul. "Block 3, Part 5." Block 3B Research Design. Walton Hall: The Open University Press, 1979.

Beck, Leonard. "Qualitative and Quantitative Conceptions of Competence: Changing Interpretations of Achievement Test Results Since 1845." (mimeographed).

Bereiter, Carl. Must We Educate?

Berrson, Robert D. "The Use of Participant Observation in the Evaluation of Art Programs." Studies in Art Education, 19, No. 2 (1978), 60-67.

Brownell, William A. "The Evaluation of Learning Under Dissimilar Systems of Instruction." (Obtained from Dr. J. Weiss.)

Cronbach, Lee J. "Course Improvement Through Evaluation." Teachers College Record, 64 (1963), 675-683.

Eisner, Elliott W. "The Perceptive Eye: Toward the Reformation of Educational Evaluation." (Invited Address) Curriculum and Objectives American Educational Research Association. Washington, D.C.: March 31, 1975.

EPIE Institute. "EPIE Materials Report No. 92m. Deciphering L*V*R*. A Key for Consumers and Publishers to Learner Verification and Revision." EPIE Institute, Box 620, Stony Brook, NY, 11790.

Glaser, Berney G. and Anselm L. Strauss. The Discovery of Grounded Theory - Strategies for Qualitative Research. Chicago: Aldine Publishing Co., 1967.

Greene, Maxine. "Curriculum and Consciousness." Teachers College Record, 73, No. 2, 253-69.

Hall, Budd L. "Participant Research: An Approach for Change." Convergence, 8, No. 2 (1975), 24-32.

Hamilton, Davis. Curriculum Evaluation. Curriculum Studies Series, ed. David Jenkins. London: Open Books, 1976.

Hanley, Janet P. et. al. "Curiosity, Competence, Community." Curriculum Design. ed. Michael Golby, Jane Greenwald and Ruth West. Toronto: John Wiley and Sons.

Hooper, Richard, ed. The Curriculum Context, Design and Development. London: Open University Press, 1971.

Jones, Richard. Fantasy and Feeling in Education. New York: NY University Press, 1968.

Kohlberg, Lawrence, and Rochelle Mayer. "Development as the Aim of Education." Harvard Educational Review, 42, No. 4 1972.

Lewy, Arieh, ed. Handbook of Curriculum Evaluation. International Institute for Educational Planning, UNESCO, 1977.

Magoon, A. Jon. "Constructivist Approaches in Educational Research." Review of Educational Research, 47, No. 4 (1977), 651-693.

Mason, Ward S. Problems of Measurement and the NIE Program. August 29, 1973.

Messick, Samuel. "Potential Uses of Noncognitive Measurement in Education." Journal of Educational Psychology. 71, No. 3, 281-292.

Oliver, Albert I. Curriculum Improvement. 2nd ed. Harper and Row, 1977.

Pace, Robert, ed. Evaluating Learning and Teaching. No. 4 (winter 1973), New Directions for Higher Education.

Pagano, Jo Anne and Lawrence Dolan. "Foundations for a Unified Approach to Evaluation Research." Curriculum Inquiry 10, No. 4 (1980), 367-380.

Palonsky, Stuart B. "Teacher Effectiveness in Secondary Schools: An Ethnographic Approach." The High School Journal. 61, No. 2 (1977). 45-51.

Parlett, Malcolm. "The New Evaluation." Trends in Education. 34, (1974) 13-18.

_____, and David Hamilton. "Evaluation as Illumination: A New Approach to the Study of Innovatory Programs." Evaluation Studies Review Annual. 140-157.

Pilsworth, Michael and Ralph Ruddock. "Some Criticisms of Survey Research Methods in Adult Education." Convergence 8, No. 2 (1975), 33-43.

Poulton, Geoff, Lawrence Randle and Paul Fordham. "A Question of Participation-action and research in the New Communities project." Convergence 8, No. 2 (1975), 54-.

Rainsberry, F. Images of Curriculum. (typescript).

Robinson, Floyd, ICPOCMU: A Framework for the Preparation of Curriculum Guidelines and Related Instructional Materials. (typescript).

Russell, Howard, ed. Evaluation of Computer Assisted Instruction Program Educational Development Innovation Diffusion. September, 1969.

Schaffarick, Jon and David H. Hampson, editors. Strategies for Curriculum Development. Berkeley, California: McCutchan, 1975.

Scriven, Michael. "The Methodology of Evaluation," Educational Evaluation: Theory and Practice, ed. Worthen and Sanders. (Worthington, Ohio): Charles A. Jones.

Smith, L.M. and Patricia C. Carpenter. General Reinforcement Package Project: Qualitative Observation and Interpretation. (St. Ann, Missouri): January 1972.

_____ and Paul A. Pohland. "Chapter 2 - Participant Observation," Evaluation of Computer Assisted Instruction Program. ed. H. Russell. September, 1969.

Stake, Robert E. "Program Evaluation, Particularly Responsive Evaluation," Paper No. 5 in Occasional Paper Series (Kalamazoo, Michigan: College of Education. Western Michigan University) November 1975.

Swantz, Marja Lusa. "Research as an Educational Tool for Development." Convergence. 8: 44-53.

Tyler, Ralph. "Purposes and Processes of Evaluation."

Vallance, Elizabeth. "The Landscape of 'The Great Plains Experience'." Curriculum Inquiry. 7 No. 2 (1977), 87-105.

Uemura, Joseph. "Measuring Man." Crucial Issues in Education. ed. Ehler.

Weiss, Joel. Affect, Affect, of Course We Have Affect— But. Speech.

Wells, Hively et al. Domain-referenced Curriculum Evaluation. CSE Monograph Series in Evaluation. Los Angeles. Center for the Study of Evaluation. 1973.

West, Gordon. "Participant Observation in Canadian Classrooms. The Need, Rationale, Technique and Development Implications." Canadian Journal of Education, 2 No. 3 (1977), 55-74.

Wheeler, D. K. Curriculum Process. Unibooks. University of London Press. 1967.

Willis, George. "Qualitative Evaluation as the Aesthetic, Personal and Political Dimensions of Curriculum Criticism." Concepts and Cases in Curriculum Criticism. Berkeley, California. McCutchan Publishing.

_____, ed. Qualitative Evaluation. includes:
"Patterns of Phenomenological Responses to Curricular Implications." G. Willis and A.J. Allen,
"Curricular Criticism and Literary Criticism" G. Willis,
"Curriculum Criticism and Literary Criticism - Comments on the Analogy" Edward F. Kelley,
"Toward a Humane Curriculum." A. W. Foshay.